IMM-NYU 332
OCTOBER 1964

NEW YORK UNIVERSITY
COURANT INSTITUTE OF
MATHEMATICAL SCIENCES

The Anomalous Concept of Statistical Evidence: Axioms, Interpretations, and Elementary Exposition

ALLAN BIRNBAUM

PREPARED UNDER
CONTRACT NO. NONR-285(38)
OFFICE OF NAVAL RESEARCH

IMM-NYU 332
October 1964

New York University

Courant Institute of Mathematical Sciences

THE ANOMALOUS CONCEPT OF STATISTICAL EVIDENCE: AXIOMS, INTERPRETATIONS, AND ELEMENTARY EXPOSITION

Allan Birnbaum

Invited paper presented at the Joint European
Conference of Statistical Societies,
Berne, Switzerland, September 14, 1964.

This report represents results obtained at the Courant Institute of Mathematical Sciences, New York University, under the sponsorship of the Office of Naval Research, Contract No. Nonr-285(38).

0. Introduction and Summary.

This paper presents some new mathematical and interpretive material on concepts of statistical evidence. The material has been given a self-contained elementary expository form (restricted to the case of discrete probability distributions) suitable for early inclusion in any mathematical statistics course however elementary. Part I (Sections 1-6) in particular is a brief but rounded elementary account of "the concept of statistical evidence and the anomalous problem of its interpretation." Part II presents further axioms for evidence and derivations of their interrelations in elementary form; its Sections 7-9 present concepts which have figured significantly in the development of statistical thinking, and are a basis for a reading of Part III. The latter contains general discussion intended to give even serious beginning students, among others, some suggestions for perspectives on the broad field of statistical theories and the historical pattern of their development.

The axioms of statistical evidence and some related concepts discussed, with their mathematical interrelations, are summarized schematically in Figure 1, p. 17a. The conceptual issues and historical patterns discussed are indicated schematically in the Figure 2, p. 33a.

PART I. The Concept of Statistical Evidence and the Anomalous Problem of its Interpretation.

1. Models of Statistical Evidence. Let E denote any specified model of a (discrete) statistical experiment, in which the random outcome X takes values in the sample space S containing points $x_1, x_2, \ldots,$ with (elementary) probability function

$$f(x,\theta) = \text{Prob}\ (X = x|\theta)\ , \qquad x \in S\ ,$$

where the parameter point θ lies in the parameter space Ω. In case both S and Ω are finite, the model may be represented conveniently by a stochastic matrix:

$$E = (p_{ij}) = \begin{pmatrix} p_{11} & \cdots & p_{1J} \\ \vdots & & \vdots \\ p_{I1} & \cdots & p_{IJ} \end{pmatrix}$$

where p_{ij} = Prob $(X=j \mid \theta=i)$, $j = 1,\ldots,J$, $i = 1,\ldots,I$. An example is the experiment

$$E_1 = \begin{pmatrix} .9 & .1 \\ .3 & .7 \end{pmatrix} .$$

Any instance of statistical evidence is represented by a model of the form (E,x), where E is a specified experiment and x is a specified observed outcome of E. An example is

$$(E_1, 2) \qquad \text{or} \qquad \left(\begin{pmatrix} .9 & .1 \\ .3 & .7 \end{pmatrix} , 2\right) ,$$

which represents that $X = 2$ was observed as the outcome of the experiment E_1.

The properties, concepts, interpretations, and uses appropriate to statistical evidence in various cases have often been found obscure and controversial. Various intuitive and mathematical concepts and objective criteria have been proposed toward adequate characterization and interpretation of statistical evidence. To discuss these, it is sometimes convenient to write Ev(E,x) to denote a concept or interpretation of the statistical evidence (E,x) or the "evidential meaning" of (E,x). Among these concepts, the simplest and intuitively most appealing seem to have a negative character, referring to aspects of evidence describable as "irrelevant," "uninformative," or "like 'noise'" (in the communication theory sense).

2. A Concept of "Irrelevant Noise," (N).

Consider any discrete experiment E, with probability function $f(x,\theta)$, $x \in S$, $\theta \in \Omega$. Let x' be a specified possible outcome of E. Let Z be an auxiliary random variable (independent of X) taking the values 1, 0, with known respective probabilities c, 1-c. Let Y be the random variable defined by

$$y = y(x,z) = \begin{cases} 1, \text{ if } x = x' \text{ and } z = 1, \\ 0, \text{ otherwise.} \end{cases}$$

Let E^* be the experiment whose outcomes are just the observed values y of the random variable Y. (E^* may be described as a

"stochastically censored" version of E. E^* has p.d.f. $f^*(y,\theta)$, $\theta \in \Omega$, given by $f^*(1,\theta) = c\ f(x',\theta)$, $f^*(0,\theta) = 1-c\ f(x',\theta)$; thus E^* is characterized fully by the probability c and the function $f(x',\theta)$, $\theta \in \Omega$. Although an experiment of the form E^* need not be considered in relation to the experiment E, it is possible to realize E^* by use of E as indicated.

Let us consider $Ev(E^*,1)$, when E^* is considered as having been actually realized by use of E. Upon observing the outcome $Y = y(x,z) = 1$ of E^* one can immediately deduce, from the definition of $y(x,z)$, that outcome x' of E has occurred; and thus one is in the same position as an experimenter who has carried out E and observed x', except in the hypothetical respect that, if x' had not occurred in E, one would in general have obtained only incomplete information concerning the outcome x of E. Many statisticians find it in accord with their concepts of evidence to consider this hypothetical distinction irrelevant; and they consequently consider $Ev(E^*,1\ )$ and $Ev(E,x')$ as equivalent in such cases.

The concept that such hypothetical distinctions are irrelevant to the evidence in question may be expressed informally as "irrelevance of stochastic censoring which might have, but in fact did not, affect an observed outcome." This concept may be formulated as:

(N): Axiom of irrelevant noise. Let E be any (discrete) experiment, with probability function $f(x,\theta)$, $\theta \in \Omega$. Let Z be an auxiliary random variable independent of X, taking values 1, 0, with respective known probabilities c, 1-c (independent

of θ). Let x' be any specified possible outomce of E. Let Y be the random variable defined by $y = y(x,z) = 1$ if $x = x'$ and $z = 1$, and $y = 0$ otherwise. Let E^* be the experiment in which Y is observed (but not X nor Z). Then $Ev(E^*,1) = Ev(E,x')$.

An example of the simplest sort is:

$$E_1 = \begin{pmatrix} .9 & .1 \\ .3 & .7 \end{pmatrix}, \quad c = \frac{2}{3}, \quad x' = 1,$$

from which we determine

$$E^* = \begin{pmatrix} .6 & .4 \\ .2 & .8 \end{pmatrix}.$$

Here (N) implies $Ev(E_1,1) = Ev(E^*,1)$.

3. The Likelihood Axiom, (L).

Another concept of statistical evidence, whose intuitive and historical background is discussed in later sections, is expressed in:

(L): The likelihood axiom: Let (E,x'), (E^*,y') be any instances of statistical evidence with common parameter space Ω, and with probability functions such that for some positive c, $f(x',\theta) = c\, f^*(y',\theta)$, $\theta \in \Omega$. Then $Ev(E,x') = Ev(E^*,y')$.

Lemma. (N) is equivalent to (L).

Proof. Under the conditions in the statement of (N), the probability function of outcome $Y = 1$ of E^* is given by $f^*(1,\theta) = c\ f(x',\theta)$, $\theta \in \Omega$. But the last relation is the condition in (L). Hence when (L) is assumed we have $Ev(E,x') = Ev(E^*,1)$, which is the conclusion of (N).

To prove the converse, as in the condition of (L) let (E_1,x_1'), (E_2,x_2'), be any two instances of statistical evidence, with common parameter space Ω and with probability functions $f_1(x_1,\theta)$, $f_2(x_2,\theta)$, such that for some positive c, $f_1(x_1',\theta) = c\ f_2(x_2',\theta)$, $\theta \in \Omega$. Without loss of generality we may assume that $c \leq 1$ (since otherwise we could write $f_2(x_2',\theta) = (1/c)f_1(x_1',\theta)$, where $(1/c) < 1$). Let E_1^* be the "stochastically censored" version of E_1, determined by taking the specified outcome x_1' and the specified probability $c_1 = 1$; thus E_1^* is characterized by $f_1^*(1,\theta) = f_1(x_1',\theta)$, $\theta \in \Omega$. Similarly let E_2^* be determined from E_2, x_2', and $c_2 = c$; E_2^* is characterized by $f_2^*(1,\theta) = c\ f_2(x_2',\theta)$, $\theta \in \Omega$. We observe that $f_1^*(1,\theta) = f_2^*(1,\theta)$, since by assumption $f_1(x_1',\theta) = c\ f_2(x_2',\theta)$, $\theta \in \Omega$. Thus $(E_1^*,1)$ and $(E_2^*,1)$ are mathematically identical, whence $Ev(E_1^*,1) = Ev(E_2^*,1)$. Further, by (N) we have $Ev(E_1^*,1) = Ev(E_1,x_1')$ and $Ev(E_2^*,1) = Ev(E_2,x_2')$. Thus $Ev(E_1,x_1') = Ev(E_2,x_2')$, which is the conclusion of (L), completing the proof.

This concept of "irrelevant noise" and its implications was introduced by the present writer (1961, pp. 418-9, 430), with examples in terms of communication channels.

Axiom (L) may be conveniently stated in terms of the <u>likelihood ratio function</u> $\lambda(\theta_1,\theta_2;x) = f(x,\theta_2)/f(x,\theta_1)$, θ_1, $\theta_2 \in \Omega$, determined by any instance (E,x) of statistical evidence: (L) is the assertion that if (E,x') and (E^*,y') determine the same likelihood ratio function, then $Ev(E,x') = Ev(E^*,y')$. The likelihood ratio function is somewhat redundant and is represented more conveniently for most theoretical and practical purposes by fixing θ_1 at any value for which $f(x,\theta_1)$ is positive and finite, or more generally by replacing in λ the factor $1/f(x,\theta_1)$ by an arbitrary positive constant c; the resulting function $c\ f(x,\theta)$, $\theta \in \Omega$, is called (a representation of) the <u>likelihood function</u>.

4. <u>Interpretations of Likelihood Functions in the Binary Case</u>.

In the case of binary experiments, that is, experiments with a parameter space of just two points, θ or i = 1,2, the likelihood function is conveniently represented by the <u>likelihood ratio statistic</u> $\lambda = \lambda(x) = f(x,2)/f(x,1)$, $x \in S$. For example in the experiment $E_1 = \begin{pmatrix} .9 & .1 \\ .3 & .7 \end{pmatrix}$, the possible outcomes j = 1, 2, determine respectively the likelihood ratios $\lambda(1) = .9/.3 = 3$ and $\lambda(2) = .1/.7 = 1/7$. If the likelihood axiom (L) is accepted as a characterization of the evidential meaning Ev(E,x) in any binary case, then Ev(E,x) is characterized fully by the number $\lambda(x)$, without other reference to the form of (E,x). The problem of interpretation of such numbers as representations of instances of statistical evidence has been discussed in detail by the present

writer (1961, 1962) and others cited there.

All considerations seem to support not only the high plausibility but the clear appropriateness of a single mode of interpretation of likelihood functions in the binary case, namely, interpretations in which numerical likelihood ratios as such are taken as indices of evidential meaning; and in which strength of evidential support for $i = 2$ as against $i = 1$ increases as λ takes higher values on the continuous scale from 0 to ∞ ; with $\lambda = 1$ representing neutral evidence equivalent to no evidence; and $\lambda = \infty$ (or 0) representing evidence of the greatest possible strength (virtually comparable to the strength of deductive logical evidence) for $i = 2$ as against $i = 1$ (or vice versa).

4.1. Error-Probabilities and the Concept of Unbiased Evidential Interpretations, (U).

One of these considerations (supporting or confirming the adequacy of this mode of evidential interpretations) is based upon the concept of error-probabilities. It seems to many, including the present writer, that one very appropriate minimum requirement for the adequacy of any mode of characterizing and interpreting statistical evidence is a requirement in terms of error-probabilities, which may be expressed as:

(U): Unbiasedness criterion for a mode of evidential interpretations:

Systematically misleading or inappropriate interpretations shall be impossible; that is, under no θ' shall there be high probability of outcomes interpreted as "strong evidence against θ'."

(U) differs in several respects from the *axioms* for statistical evidence formulated above and below. Each axiom specifies the equivalence of certain instances of evidential meanings and interpretations, without referring to the form or nature of the latter concept. (U) does refer to these, with a measure of vagueness which is necessary at least at this stage of the present discussion, but with sufficient precision to allow clear demonstration that the criterion (U) is met very well by (L) and the mode of interpreting λ values indicated above.

The essential reason for this adequacy is seen clearly in the very definition of $\lambda(x) = f(x,2)/f(x,1)$: For λ values which are much larger than unity (e.g. $\lambda \geq 100$), which would be interpreted as relatively strong evidence for $i = 2$ as against $i = 1$, the probability in any experiment whatever is at least 100 times as large under $i = 2$ (the case in which evidence thus interpreted is highly appropriate and desirable) as under $i = 1$ (the case in which evidence thus interpreted is highly inappropriate and misleading. A parallel comment applies to λ values much smaller than unity (e.g. $\lambda \leq .01$). And λ values not far from unity, which would be interpreted as weak evidence as between $i = 1$ and $i = 2$, have, in any experiment whatever, probabilities which are not far from unity in ratio; for example, the probability of $\lambda = 1$, which would be interpreted as strictly neutral or uninformative evidence, is the same under $i = 1$ and $i = 2$ in every experiment. (U) is satisfied since the definition of $\lambda(x)$ leads directly to bounds such as $\Pr(\lambda \leq .01|i=2) \leq (.01) \Pr(\lambda \leq .01|i=1) \leq .01$. Of course experiments in which weak evidence has small probabilities

are preferable. Further detailed examples and interpretations were given by the writer (1961, 1962).

Not only does (L), together with such evidential interpretations of λ values, meet the condition of adequacy suggested by the general error-probability concept, but it does so in a manner which is free of the obscure, awkward, or implausible features found in alternative, more standard modes of evidential interpretations:

In the Neyman-Pearson approach, which takes error-probabilities as its single basic concept, the requirement to fix a Type I error-probability in order to determine a Type II error-probability forces upon evidence the unnatural dichotomy into "rejection" and other. In fact common practice modifiesor ignores this dichotomization in favor of the more plausible reporting of P-levels lying in the range of possible Type I error-probabilities, typically without reference to associated Type II error-probabilities. In neither case does the standard testing approach include any definite concepts of evidential interpretation associated with error-probabilities. Neyman himself has been foremost in insisting that standard statistical methods such as estimation and testing cannot appropriately be interpreted as methods of inference in the sense of evidential interpretations (e.g. 1957, 1962). Nevertheless the latter interpretation is taken as basic, implicitly if not explicitly, in a major part of exposition and application of the Neyman-Pearson theory. Such typical interpretations are made clearly explicit in, for example, the modern textbook by Walker and Lev (1953, p. 54) where an interpreted term distinct from

probability, called "confidence", is introduced to represent the evidential interpretation of a typical estimate given by the Neyman-Pearson theory: "To distinguish a confidence statement from a probability statement, this text will use the notation Conf$(.05 \leq P \leq .65) = .95$ . ... This may be expressed in words thus: "We have confidence .95 that the unknown proportion lies between .05 and .65" or "We assert that P is not smaller than .05 and not larger than .65 and we have arrived at these numbers by a procedure which if applied repeatedly would yield interval estimates of which 5% would not contain the true value..." The presentation as alternatives of these two interpretations of the new term also makes explicit the widespread view that the basic error-probability property stated last is tantamount to or at least warrants an evidential interpretation involving concepts other than error-probabilities.

Finally, we note that the approach based exclusively upon error-probabilities lacks even a basis for confronting such plausible concepts of evidence as that of "irrelevant noise," (N), introduced above (or the weaker "sufficiency" concept (S) discussed below).

5. Interpretations of likelihood functions in the general case.

The only mode which has been suggested for evidential interpretation of likelihood functions in general is due to Fisher (e.g.1956) and Barnard (e.g.1962).This includes the interpretations described above for the binary case, where it appeared eminently satisfactory. But in the general case it appears crucially incomplete and

inadequate in several respects.

Consider the example of a single observation on X, assumed normally distributed with unknown mean μ and standard deviation σ, $-\infty < \mu < \infty$, $\sigma > 0$. Any observed value x determines a likelihood function (represented, up to an arbitrary positive constant factor, by)

$$L(x,\mu,\sigma) = \frac{1}{\sigma} \exp - \frac{1}{2\sigma^2} (\mu-x)^2 , \quad - \infty < \mu < \infty , \sigma > 0.$$

For each arbitrarily large finite number M, and each arbitrarily small positive $m < M$, $L(x,\mu,\sigma) \geq M$ occurs only on (a subset of the) parameter points for which

$$\sigma \leq K_M \qquad \text{and} \qquad x - K_M' \leq \mu \leq x + K_M' ,$$

where K_M, K_M' are numbers depending just on M; and $L(x,\mu,\sigma) \leq m$ occurs on every parameter point for which $\sigma \geq k_m$, where k_m depends just on m. The ratio K_M/k_m can be made arbitrarily large by suitable choice of M and m. All considerations toward a plausible mode of evidential interpretation of likelihood functions in general seem to suggest that in such a case small values $\sigma \leq k_m$ are to be considered as supported by the observed statistical evidence, as against large values $\sigma \geq K_M$ (possibly with addition of considerations linking certain μ values to certain σ values), with a strength which is very high with suitable values of M and m. However consider any fixed M, m, μ' and $\sigma' > K_M$; for this parameter point the probability is unity of an observation x and a likelihood function L of a form leading to the evidential interpretations just described, which must evidently be considered strongly

misleading and inappropriate when this parameter point is true.

(All interpreted features of this example could be duplicated in a discrete example, e.g. a family of discrete distributions of X closely approximating the respective normal distributions of the example.)

Thus in the general case, unlike the binary case, the sole suggested (Fisher-Barnard) mode of evidential interpretations compatible with (L) fails grossly to satisfy the adequacy condition (U) suggested by the Neyman-Pearson concept of error-probabilities. Analogous conclusions were supported by Neyman (1938; 2nd. ed., 1952), Armitage (in Smith (1961), pp. 32-34), and Stein (1962) by consideration of experiments of more complicated forms (asymptotic or sequential).

(The concepts of intrinsic confidence and significance levels discussed by the present writer (1962, Part II) seem to have at most heuristic, but not substantial, value for evidential interpretations. We exclude from the present discussion the Bayesian approach, in which a characterization of evidence formally like (L) is deduced from other concepts, but in which it may be said that no autonomous role is played by a concept of empirical statistical evidence. Some difficulties of interpretation of likelihood functions are to be avoided, according to Fisher and Barnard, by placing limitations on the scope of (L) and by alternative use of the approach of fiducial probability. But to many, including the present writer, such limitations and alternatives seem beset with even greater difficulties and obscurities than those besetting the likelihood approach, and furthermore to

be incompatible with the likelihood concept rather than complementary to it.)

6. The Anomaly of the Empirical Concept of Statistical Evidence.

Empirical evidence generally and statistical evidence in particular are integral parts of the structure, practice, and process of science. This fact is represented from the standpoint of the working scientific research investigator by Wilson (1952), for example; from the standpoint of broad logical and philosophical analysis of the structure of science, by Nagel (1961), for example; and from the standpoint of mathematical statistics, by a major portion of its literature, both basic and expository, and of general practice of its applications. Throughout much of this braod body of thought and practice there has run a seriously-held though often tacit assumption that there exist, actually or at least potentially, concepts of empirical evidence and of statistical evidence adequately clear and appropriate to account for the nature and importance of evidence in the structure and process of science. Concerning empirical statistical evidence in particular, undoubtedly one of the strongest forces in the development of mathematical statistics has been its accepted specialized but significant responsibility for clarifying basic concepts (as well as developing working techniques) of empirical statistical evidence, as such and in relation to other factors in the process of scientific work.

In connection with the more general concept of empirical evidence, it is seen in works like those of Wilson and Nagel

cited and in the general practice of scientific work that there do not now exist very precise concepts of the nature and role of empirical evidence in science; for example, no very precise account can be given of the articulation of empirical evidence with the several other ingredients in the structure of an empirical law or theory. And much experience and thought leaves unsupported the view that clear and precise concepts of empirical evidence do exist somehow implicitly or tacitly in science or that they are likely to be discovered or developed.

Concerning the more specific problem of developing a precise and adequate empirical concept of statistical evidence, the preceding sections give an elementary and self-contained demonstration that this problem provides an intriguing and significant mystery: Notwithstanding the evident fact that statistical evidence is very widely and effectively interpreted and used in scientific work, no precise adequate concept of statistical evidence exists -- and <u>none can exist</u>, in the sense that precise versions of several very plausible widely-accepted minimum conditions for adequacy of such a concept are logically incompatible! The mathematician's natural response to a disclosed contradiction, that of selecting consistent subsets of conditions on which to develop consistent theories, is not available here. So long as the incompatible criteria appear to be appropriate mathematical expressions of respective aspects of an extra-mathematical entity (in this case the more-or-less coherent, more-or-less shared, more-or-less explicit body of concepts of empirical statistical evidence which are a part of the structure and process

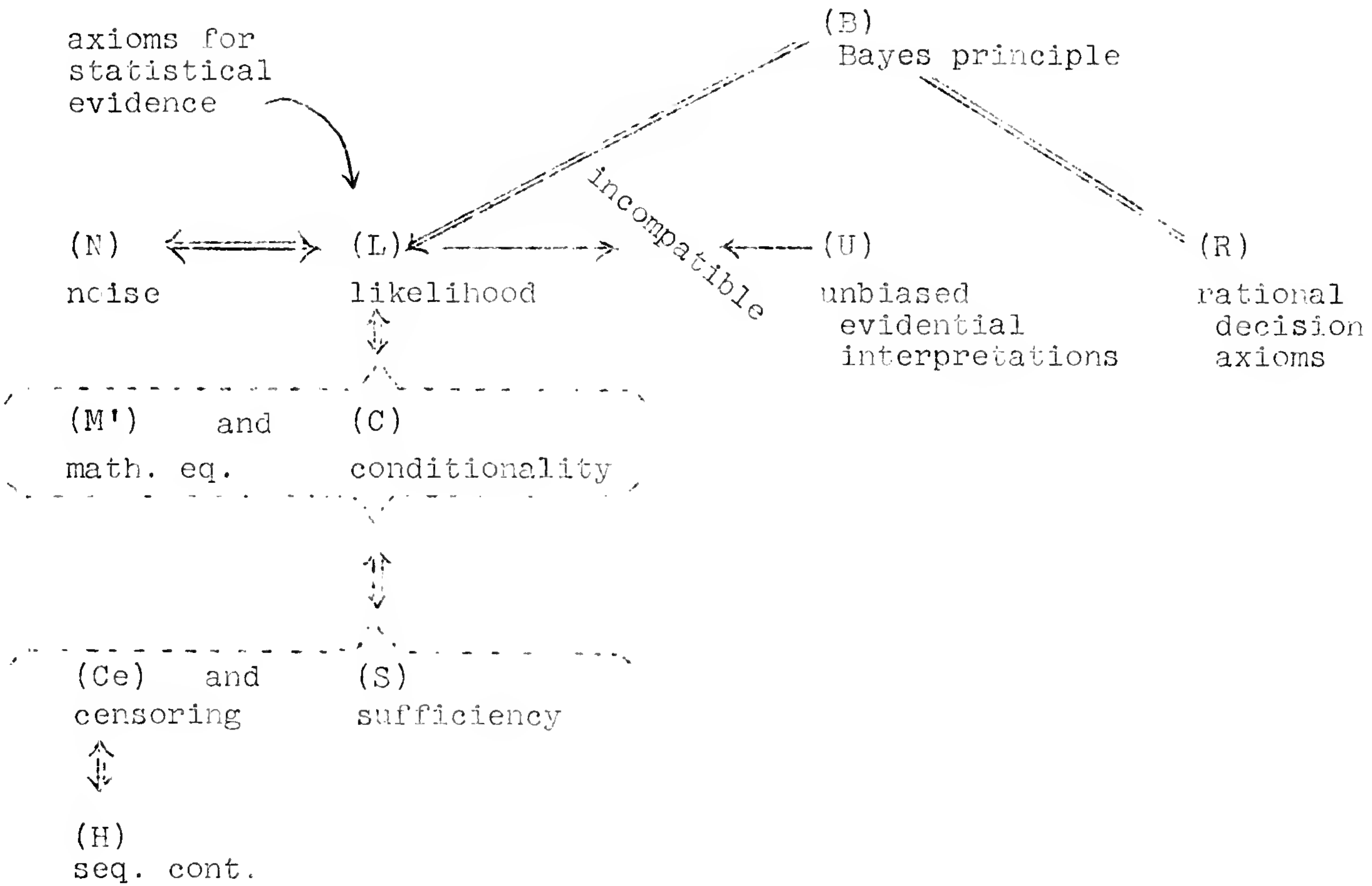
of science), so long does the incompatibility express an evident anomaly in that entity. In the writer's opinion this evident anomaly is a substantial one. It seems worthy of broad serious intellectual curiosity; and particularly for those working in theoretical and applied statistics, consideration of it can be a salutory antidote for over-facile views on the relations among concepts and between concepts and scientific practice.

This anomaly, which is the evident outcome of the problem of an adequate precise concept of statistical evidence, is a vantage-point for an interesting and orderly perspective on the pattern of historical development of theories of statistical inference, since in most periods the strongest impetus for new developments or changes has been the specific inadequacies of current concepts and techniques for treating statistical evidence. Such a perspective is presented briefly in Section 13 below, and also summarized graphically in Figure 2, p. 33a.

PART II. Further Axioms for Statistical Evidence.

The following Sections 7-11 complement Part I by introducing in elementary mathematical form further axioms for statistical evidence, with derivations of all implications between axioms, as summarized in Figure 1, p. 17a. Of these, Sections 7-9 are an adequate preparation for reading the discussion of Part III.

Figure 1. Axioms for statistical evidence, and related concepts, with their logical interrelations.



7. Mathematical Equivalence, (M').

If two experiments differ only in the manner of labeling of sample points, (i.e. only by a one-to-one transformation of sample space preserving probabilities), they are usually taken as equivalent for all purposes. For example

$$E_2 = \begin{pmatrix} .1 & .9 \\ .7 & .3 \end{pmatrix}$$

is thus equivalent to E_1 above. For this reason, corresponding outcomes of such equivalent experiments are usually considered as providing equivalent statistical evidence. For example $(E_2,1)$ is thus equivalent to $(E_1,2)$. This concept may be expressed informally as "irrelevance of manner of labeling outcomes," and suggests the simplest and weakest among the several axioms for statistical evidence to be considered here. It suffices for our purpose to consider an axiom expressing only a special small part of the concept of equivalence just indicated:

(M'): Axiom of mathematical equivalence: If x' and x'' are possible outcomes of any discrete experiment E, with identical probability functions $f(x',\theta) = f(x'',\theta)$, $\theta \in \Omega$, then $Ev(E,x') = Ev(E,x'')$.

For example, under (M') we have

$$Ev\left(\begin{pmatrix} .1 & .1 & .8 \\ .4 & .4 & .2 \end{pmatrix}, 1\right) = Ev\left(\begin{pmatrix} .1 & .1 & .8 \\ .4 & .4 & .2 \end{pmatrix}, 2\right) .$$

8. A Weak Concept of "Irrelevant Noise": the Sufficiency Axiom, (S).

Let E be any specified discrete experiment, with sample space S and probability function $f(x,\theta)$, $\theta \in \Omega$. Let E be augmented as follows: if and when a certain outcome x' of E is observed, an observation y is taken on a random variable Y which has possible outcomes 1, 2, with respective known probabilities c, 1-c. When any outcome of E other than x' is observed no auxiliary observation is taken. The augmented experiment is an experiment E^* with sample space

$$S^* = \{x \mid x \in S,\ x \neq x',\ \text{or } x = (x',1) \text{ or } (x',2)\} ,$$

and probability function

$$g(x,\theta) = \begin{cases} f(x,\theta), & x \in S,\ x \neq x' , \\ cf(x',\theta), & x = (x',1) , \\ (1-c)\ f(x',\theta), & x = (x',2) . \end{cases}$$

Since the auxiliary random variable Y has a known distribution independent of θ, an observed value such as y is considered by many statisticians as representing only recognizable added "noise," irrelevant to θ and thus irrelevant to $Ev(E^*,(x',y))$; and thus $Ev(E^*,(x',y))$ is considered equivalent to $Ev(E,x')$. This concept of "irrelevance of recognizable pure noise" for statistical evidence may be formulated as:

(S): Sufficiency axiom: Let E and E^* be any discrete experiments with common parameter space Ω. Let one possible outcome x' of E have probabilities $f(x',\theta)$; let two outcomes (x',1),

$(x',2)$ of E^* have respective probabilities $cf(x',\theta)$, $(1-c)f(x',\theta)$, $\theta \in \Omega$, with c known. Let the remaining possible outcomes be in one-to-one correspondence with common labels x $(x \neq x')$ and common respective probabilities $f(x,\theta)$. Then $\mathrm{Ev}(E^*,(x',1)) = \mathrm{Ev}(E,x')$.

An example of (S) of the simplest sort is:

$$\mathrm{Ev}\left(\begin{pmatrix} .9 & .1 \\ .3 & .7 \end{pmatrix}, 1\right) = \mathrm{Ev}\left(\begin{pmatrix} .6 & .3 & .1 \\ .2 & .1 & .7 \end{pmatrix}, 1\right) .$$

(Under the conditions of (S), the statistic x in E^* provides the simplest sort of example of a <u>sufficient statistic</u>.)

Lemma. (S) implies (M').

The proof is immediate upon taking $c = 1/2$.

9. The Conditionality Axiom, (C).

Let E_1, E_2 be any discrete experiments with common parameter space Ω, with respective sample spaces $S_1 = \{x_1\}$, $S_2 = \{x_2\}$, and respective probability functions $f_1(x_1,\theta)$, $f_2(x_2,\theta)$. Let E^* be the "mixture" experiment defined as follows: an auxiliary random variable Y with outcomes 1, 2, with respective known probabilities c, 1-c (not depending upon θ) is observed. If $Y = 1$ is observed, then E_1 is carried out and some outcome x_1 is observed; if $Y = 2$, then E_2 is carried out and some x_2 observed. Thus outcomes of E^* have the form (E_1,x_1), $x_1 \in S_1$, or (E_2,x_2), $x_2 \in S_2$; and respective probabilities $cf_1(x_1,\theta)$ and $(1-c)f_2(x_2,\theta)$.

The notion that $Ev(E^*,(E_1,x_1))$ is the same as $Ev(E_1,x_1)$ may be described informally as "irrelevance of experiments which might have been, but were not, carried out" and as "appropriateness of conditional interpretations of evidence." This conditionality concept has been discussed at length (Birnbaum (1962) and references therein); it may be formulated as:

(C): Conditionality axiom: If E^* is a mixture of two discrete experiments E_1, E_2, with given respective probabilities c, 1-c (independent of θ), then for any x_1', $Ev(E^*,(E_1,x_1')) = Ev(E_1,x_1')$.

An example of (C) of the simplest sort is:

$$Ev\left(\begin{bmatrix} .9 & .1 \\ .3 & .7 \end{bmatrix}, 1\right) = Ev\left(\left[\frac{1}{2}\begin{bmatrix} .9 & .1 \\ .3 & .7 \end{bmatrix}, \frac{1}{2}\begin{bmatrix} .6 & .4 \\ .4 & .6 \end{bmatrix}\right], 1\right) .$$

(It is readily verified that the 2 x 4 stochastic matrix in the right member represents the equal-weighted mixture of experiments

$$\begin{pmatrix} .9 & .1 \\ .3 & .7 \end{pmatrix} \text{ and } \begin{pmatrix} .6 & .4 \\ .4 & .6 \end{pmatrix}.)$$

Lemma. (C) and (M') jointly are equivalent to (L).

Proof. Let (E_1,x_1'), (E_2,x_2') be any instances of statistical evidence with common parameter space, with probability functions satisfying $f_1(x_1',\theta) = cf_2(x_2',\theta)$, $\theta \in \Omega$, for some positive c; without loss of generality we assume $c \leq 1$. Consider the mixture experiment E^* of E_1, E_2, with respective probabilities $1/(1+c)$, $c/(1+c)$. We have by (C) that $Ev(E^*,(E_1,x_1')) = Ev(E_1,x_1')$ and $Ev(E^*,(E_2,x_2')) = Ev(E_2,x_2')$. In E^*, outcome (E_2,x_2') has probabilities $(c/(1+c))\ f_2(x_2',\theta)$, and outcome (E_1,x_1') has probabilities $(1/(1+c))\ f_1(x_1',\theta)$; these, by the original assumption, are equal for each $\theta \in \Omega$. Hence by (M'), $Ev(E^{**},(E_1,x_1')) = Ev(E^*,(E_2,x_2'))$. Thus $Ev(E_1,x_1') = Ev(E_2,x_2')$, completing the proof that (C) and (M') imply (L).

It is obvious that (L) implies (C) and (M').

10. A Concept of Irrelevant Censoring, (Ce).

Let x' be a specific possible outcome of any discrete experiment E, with probability function f(x,Θ). Let E' be an experiment, which may be described as a "censored" version of E, in which an outcome also labeled x' has the same respective probabilities f(x',Θ) as in E, and there is just one other possible outcome x^*, for which the respective probabilities are necessarily 1-f(x',Θ). (E' is related to E by the particular many-to-one ("censoring") transformation of sample space which preserves x' and carries all other sample points of E into x^*. E' is "less informative" than E, in the sense defined in the theory of comparison of experiments.)

We now compare Ev(E,x') with Ev(E',x'): E' is equivalent to an experiment in which E is carried out, and the outcome is described only as either "x'" or "not x'." In the case of the outcome (E',x'), then, one is in exactly the position of an experimenter who has carried out E and observed x', except in the hypothetical respect that, if x' had not occurred in E_1, one would then have obtained only the incomplete information "not x'." Many statisticians find it in accord with their concepts of evidence to consider this hypothetical distinction irrelevant, and accordingly consider Ev(E',x') and Ev(E,x') equivalent in such cases. The point can be further illustrated in terms of a non-statistical example (a non-probabilistic analog of the example employed by Pratt (1961, 1962) in originating this concept and showing its consequences): If an accurate voltmeter gave a reading of 87, does it matter, for the interpretation and usefulness of

this reading (assumed error-free), whether the meter's range was bounded by 1,000, or bounded by 100? In the latter case, readings between 100 and 1,000, which might have occurred but in fact did not, would have been indistinguishable.

The concept that such hypothetical distinctions are irrelevant to the evidence in question may be expressed informally as "irrelevance of censoring which might have, but in fact did not, affect an observed outcome." Since the structure of E' is determined by just the function $f(x',\theta)$, $\theta \in \Omega$, this concept may be formulated as:

(Ce): Axiom of irrelevant censoring: For any specified outcome x' of any discrete experiment E, Ev(E,x') is characterized fully by just the function $f(x',\theta)$, $\theta \in \Omega$, without other reference to E or x'.

An example of (Ce) of the simplest sort is:

$$\mathrm{Ev}\left(\begin{bmatrix} .9 & .1 \\ .3 & .7 \end{bmatrix}, 1\right) = \mathrm{Ev}\left(\begin{bmatrix} .9 & .05 & .05 \\ .1 & .05 & .65 \end{bmatrix}, 1\right) .$$

Lemma. (Ce) implies (M').

Proof. In E, let x', x" be such that $f(x',\theta) = f(x'',\theta)$, $\theta \in \Omega$, as in the condition of (M'). Then, assuming (Ce), we have Ev(E,x') = Ev(E,x"), the conclusion of (M').

Lemma. (Ce) and (S) jointly are equivalent to (L).

Proof. Assuming (Ce) and (S), consider any (E,x') and (E^*,y') for which $f(x',\theta) = cf^*(y',\theta)$, $\theta \in \Omega$, $c > 0$. If $c = 1$, (Ce) gives that $\mathrm{Ev}(E,x') = \mathrm{Ev}(E^*,y')$. If $c \neq 1$, we may assume without loss

of generality that $0 < c < 1$. (If $c > 1$ we could write $f^*(y',\theta) = (1/c)\ f(x',\theta)$ where $0 < (1/c) < 1$.) Consider the auxiliary random variable Z with outcomes 1, 2, with respective probabilities c, 1-c; Z is to be observed only if and when outcome y' of E^* is observed. In the augmented experiment E^{**}, the outcome (y',1) has probabilities $cf^*(y',\theta)$, equal to those of x' of E; hence by (Ce) we have $Ev(E^{**},(y',1)) = Ev(E,x')$. And by (S), $Ev(E^{**},(y',1)) = Ev(E^*,y')$. Hence $Ev(E,x') = Ev(E^*,y')$, completing the deduction of (L). The converse is obvious.

11. A Concept of Irrelevance of Some Hypothetical Continuations of Sequential Experiments, (H).

Consider any two sequential (discrete) experiments E_1, E_2, with common parameter space Ω, in which the first stages have the same form. More specifically, suppose that in each experiment the first stage consists of observing a random outcome X_1 taking just the values x' or x" with respective probabilities $g(x',\theta)$, $g(x'',\theta) = 1 - g(x',\theta)$, $\theta \in \Omega$, with each experiment terminating in the first stage if and only if $X_1 = x'$ is observed. If $X_1 = x''$, both experiments continue but with different rules for further observation and termination.

Let us consider and compare $Ev(E_1,1)$ and $Ev(E_2,1)$. Suppose that two scientists independently investigating the same subject-matter, represented by the common parameter space Ω, adopt two different experimental procedures represented by E_1 and E_2. Suppose that, unknown to each other, they entrust the execution of their respective experiments to the same laboratory technician,

who notices the common form of the first stages of E_1 and E_2. Suppose also that the technician appreciates that he can economize without invalidating either experiment by taking a single observation on X_1, rather than one for E_1 and another for E_2, and does so, and happens to observe $X_1 = x'$. He properly reports to the first scientist the statistical evidence (E_1,x') and to the second scientist the formally different statistical evidence (E_2,x'), although of course the two reports are descriptions of a single physical experimental situation and outcome.

In such a case many statisticians and scientists would consider $Ev(E_1,x')$ and $Ev(E_2,x')$ equivalent, and would consider as irrelevant to the evidence in question the physically-hypothetical distinction that <u>if</u> $X_1 = x'$ had not occurred then the further physical realizations of the respective experiments would in general have taken different forms. The latter concept may be expressed informally as "irrelevance of differences between sequential sampling rules which might have, but in fact did not, affect an observed outcome."

A further illustration of this concept is provided by considering the situation of the technician in the example on the assumption that he also has an independent scientific interest in the subject matter. He must carry out a definite sequential experiment which we denote by E_3, which begins with an observation of X_1 as its first stage and which terminates there only if $X_1 = x'$. In the remaining case his experiment continues according to some definite plan as required to complete both E_1 and E_2. (In general E_3 will be more informative than either E_1 or E_2

since in general E_3 requires further observations after the termination of E_1 or E_2.) Thus in the case of outcome $X_1 = x'$ the technician obtains the statistical evidence (E_3,x') which is mathematically distinct from (E_1,x') and from (E_2,x'). Considerations like those above suggest the equivalence of $Ev(E_3,x')$, $Ev(E_1,x')$, and $Ev(E_1,x')$.

This concept may be formulated as:

(H): Axiom of irrelevance of hypothetical sequential continuations. Let the distinct (discrete) sequential experiments E_1, E_2, with common parameter space Ω, have identical first stages consisting of an observation on a random outcome X_1 taking just the values x', x'', with respective probabilities $g(x',\theta)$, $1-g(x',\theta)$, $\theta \in \Omega$. Let $X_1 = x'$ be a termination point for each experiment. Then $Ev(E_1,x') = Ev(E_2,x')$.

An example of the simplest sort is the following: Let $\theta = .01$ or $.99$ only; let $g(1,\theta) = \theta$, $g(2,\theta) = 1-\theta$. In E_1 let $X_1 = 2$ be followed by a second and final observation on a random variable X_2 independentof, and having the same distribution as, X_1. In E_2, let $X_1 = 2$ also be a termination point. (Thus E_2 is not sequential; but it can be made sequential in a formally genuine but trivial sense by adding, when $X_1 = 2$, a further observation on a random variable Y with known distribution (not depending upon θ and thus providing no information when observed).)

Lemma. (H) is equivalent to (Ce).

Proof. Clearly (Ce) implies (H), since in the latter's assumptions $(E_1,1)$ and $(E_2,1)$ have identical probabilityfunctions $g(1,\theta)$. To

prove the converse, let (E_1,x'), (E_2,x') be any instances of statistical evidence with common parameter space Ω and probability functions such that $f_1(x',\theta) = f_2(x',\theta)$, $\theta \in \Omega$. Let E_1' be the "censored" version of E_1, with outcomes $y = y(x) = x'$ if $x = x'$ and $y = 0$ otherwise; E_1' has the p.d.f.

$$g(y,\theta) = \begin{cases} f_1(x',\theta) & \text{if } y = 1 , \\ 1-f_1(x',\theta) & \text{otherwise,} \end{cases} \qquad \theta \in \Omega.$$

Let E_1'' be the "truncated" version of E, from which possible outcome x' is deleted and other outcomes have corresponding probabilities; E'' has the p.d.f.

$$h_1(x,\theta) = \frac{f_1(x,\theta)}{1-f_1(x',\theta)} , \qquad x \in S'', \quad \theta \in \Omega ,$$

where S'' denotes S with x' deleted. Let E_1^* be a two-stage sequential experiment defined as follows: (1) E_1' is carried out, and if $y = 1$ is observed the experiment E_1^* is terminated, but otherwise: (2) E_1'' is carried out. The possible outcomes z of E^* are thus denotable as $z = x'$ (representing termination in stage (1) with observation of $y = 1$), and $z = x$ for each possible value x other than x' (representing each possible termination in stage (2); thus the sample space of E_1^* is identical with that of E_1. It is readily verified that the probability function $f_1^*(z,\theta)$ of E^* is identical with the probability function $f_1(x,\theta)$. Let E_2', E_2'', and E_2^* be defined analogously in terms of E_2. Then the sequential experiments E_1^*, E_2^*, satisfy the assumptions of (H), and assuming (H) we have that $Ev(E_1^*,x') = Ev(E_2^*,x')$. By the

mathematical equivalence of (E_1,x') with (E_1^*,x'), and (E_2,x') with (E_2^*,x'), we have also $Ev(E_1,x') = Ev(E_2,x')$, the conclusion of (Ce), completing the proof.

Since (Ce) and (S) were shown above to be jointly equivalent to (L), we see that (H) and (S) are jointly equivalent to (L).

PART III. Discussion.

12. Evidence, Error-Probabilities, and Inductive Behavior (Decision-Making).

The leading theory of statistical inference, that of Neyman and Pearson, draws its strength from the fact that it bases itself rather exclusively on the highly successful modern theory of probability and its applications. In contrast with other theories of statistical inference, it avoids introducing basic interpreted terms other than error-probabilities, and gives to the latter the frequency interpretation widely accepted for probabilities generally.

The very plausible widely-held concept that a systematic connection of some kind must exist between suitably specified evidential interpretations and error-probabilities has been expressed in the present paper in the criterion (U) stated in Section 4 above. On the other hand, each of the very plausible concepts expressed in the various *axioms* formulated above implies a criticism and rejection of the dominant view that error-probabilities can serve as the sole basic term in an adequate mode of evidential interpretations. (These incompatibilities have been discussed in detail by the writer (1961, 1962) and others cited therein.)

The incompatibility in principle between evidential interpretations typically made in current practice and the most widely accepted theoretical basis for that practice is also illustrated very sharply in another way, by reference to recent advances hailed

by Neyman as "breakthroughs in the theory of statistical decision making." Neyman (1962, p. 22) writes "The essence of Robbins' second breakthrough (of 1950) is the information that, if a statistician deals simultaneously with a large number N of identical problems of testing hypotheses and wants to diminish the overall expected frequency of errors of both kinds, then, at least in certain circumstances, he can bring this frequency below the level attainable through N independent applications of the most powerful test. The resulting gain may be impressive." This quotation can be paraphrased to supply an analogous interpretation of Stein's (1955) result: If a statistician deals simultaneously with N ($N \geq 3$) identical problems of point-estimation, with independent known normal error-distributions, and wants to diminish the overall expected mean squared error, he can do so by replacing the classical estimates (the respective independent sample means) by estimates of respective means each of which in general depends on other samples besides that whose distribution is determined by the estimated mean."

(It is important to appreciate that although the parameter spaces of the respective experiments are assumed to have the same form, the subject-matter under investigation is assumed to be distinct in the respective experiments. If the subject-matter were the same, the overall experiment would simply take the form of N independent replications of one experiment, and there would be no question of compounding to be considered. Of course the assumption of common forms is made to give a simple case and is not an essential restriction in the compounding approach.)

A formally valid application of the Stein's compounding method would be to "improve" published lists of various physical constants (when measurement errors could be assumed known, normal, and independent and assigned a common scale-unit) by recomputing them in the indicated pooled or compounded way. No barrier nor qualification concerning such applications is to be found either in the formal assumptions or in Neyman's recommendations for very broad application of these methods. However an application like that described would be grossly incompatible with unformalized but firmly held concepts of many scientists and statisticians concerning the autonomy of units of experimental evidence arising in natural or experimental situations regarded as unrelated.

(An exposition of the compounding approach suitable for early inclusion in any course introducing the mathematics of decision theory is that of Robbins and Samuel (1961).)

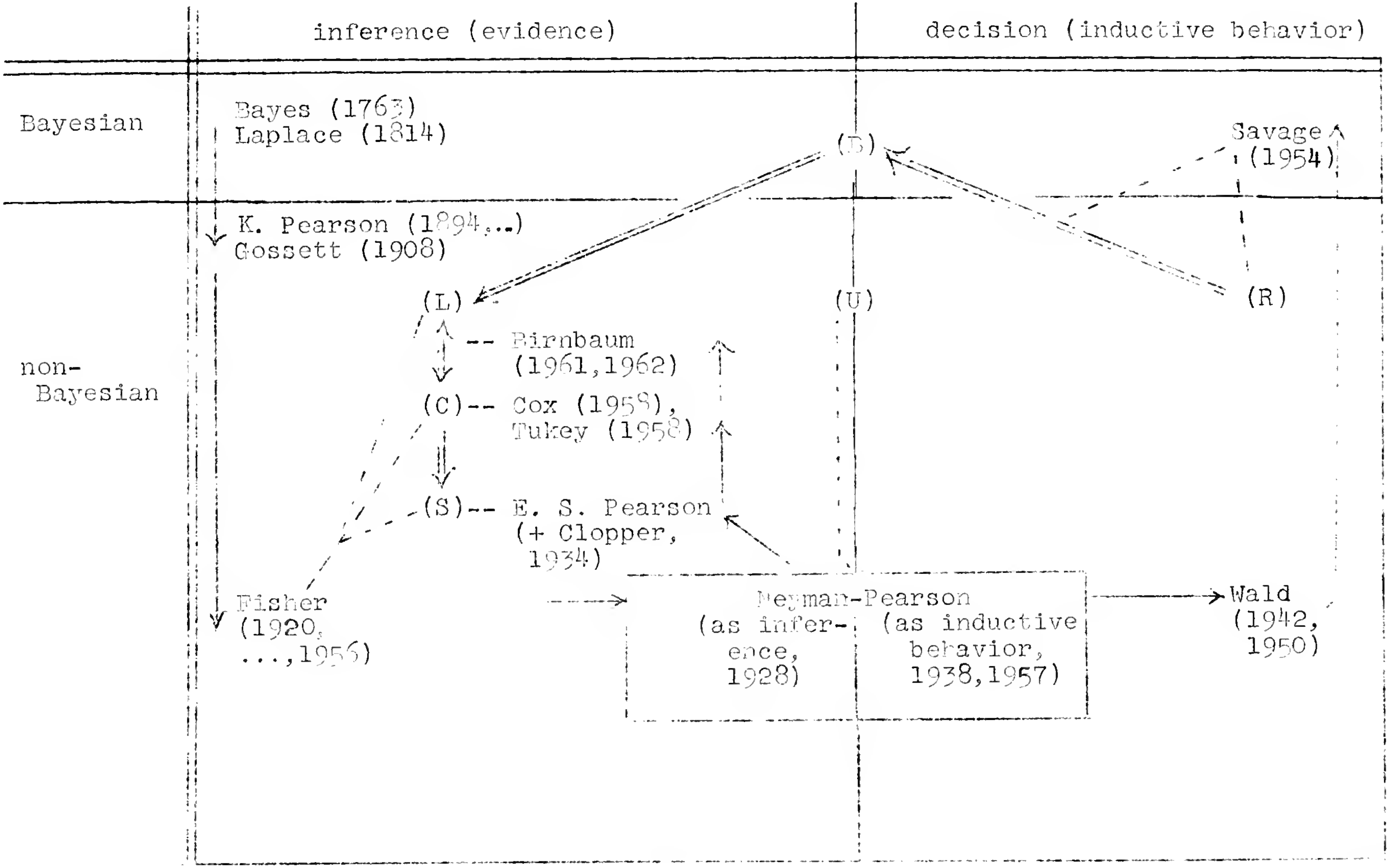
It is interesting to note that the likelihood axiom (apart from questions of interpreting likelihood functions) entails an automatic provision for a concept of autonomy between units of independent statistical evidence concerning distinct subject-matters. The notion of independent units of statistical evidence requires formalization in the product rule of probability for independent events, corresponding here to a product of density functions for respective independent samples; the notion of distinct subject-matters requires over-all formalization in a parameter space which is the cartesian product of respective-subject-matter parameter spaces; and the result is a single overall model of an experiment with the features mentioned. The product relation entails that

each likelihood function of an outcome of this experiment will have the recognizable form of a product of "marginal" likelihood functions, each representing evidence concerning one distinct subject-matter.

13. General Discussion of Concepts of Statistical Evidence.

Some of the axioms for statistical evidence formulated above, and related concepts, have sometimes been referred to previously by the present writer and others as "principles" rather than axioms. The term "axiom" has been used exclusively here to avoid possible misunderstandings of intention or possibly-misleading connotations such as may have prompted the comments of Neyman (1962, pp. 24-5). Perhaps it is worth remarking that here as elsewhere axioms have been formulated to characterize a mathematical subject-matter, and their relations and implications developed in the mathematical way, independent of possible interpretations and applications. And here as elsewhere, the adequacy of each proposed interpretation or application of an axiom or a mathematical consequence, in relation to someone's concepts or (extra-mathematical) theories or practical purposes, is a quite distinct question. In the latter connection, the subject-matter of concepts of statistical evidence has of course its own difficulties and obscurities, some of them of very long standing; if this subject-matter may be compared broadly with other subject-matters of applied mathematics, it is perhaps marked by the manner and degree in which it is conceptual rather than physical.

Figure 2. Successive developments in mathematical statistics (———>) with relations to concepts of evidence (----). (Cf. Fig. 1.)



Of course the examples and interpretations accompanying the concepts formulated above are to be taken as illustrative, and as representing the mode in which some mathematical statisticians have expressed and attempted to refine their concepts of statistical evidence by formulation of axioms. In this area, each interested person is peculiarly able to practice applied mathematics independently, and to be in principle his own ultimate authority concerning the adequacy or tenability, interpretability and usefulness of eacn axiom, each consequence, and of the entire mathematical approach to this subject-matter. Of course such independent critical consideration of basic concepts is a necessary condition for a much-to-be-desired genuine concensus on such concepts.

These concepts and axioms about evidence can usefully be regarded in each of three ways: (a) in critical confrontation with one's own concepts of evidence; (b) as concepts of some prominent theoretical statisticians, which have played a role of some significance in the development of statistical thought; and (c) as a body of mathematical material whose structure is not without interest even when considered apart from any interpretations and extra-mathematical implications.

As is familiar to teaching and consulting statisticians, students and users of the Neyman-Pearson approach and of Neyman's concept of "inductive behavior" (or decision-making as contrasted with inference) frequently meet difficulties through trying to relate their concepts of evidence to the Neyman-Pearson theory;

and frequently find such attempts difficult to give up because for many individuals concepts of evidence are among the strongest and most interesting intuitive concepts engaged by study or use of statistics. Even for those who find little value or interest in concepts of statistical evidence, a positive elementary approach to these concepts may have at least the value of helping to clarify and crystalize this direction of thought even if only to facilitate its distinction from another preferred approach.

In this spirit it seems useful and interesting to consider broadly the pattern of historical development of theories of statistical inference, particularly with reference to the impetus provided at each stage by concern for a more adequate concept of statistical evidence. Such a pattern is indicated schematically in Fig. 2, p.33a. Even after qualifications about possible additions and refinements are taken into account (these will not be discussed here), the indicated pattern seems to have substantial accuracy and interest.

The axioms of evidence discussed in the present paper are to be located in the category of non-Bayesian inference in Figure 2. They may be said to stem from some (but not all) of R. A. Fisher's theories of inference; and from the concepts of evidence of some theoretical statisticians oriented more closely to the Neyman-Pearson theory. Of course the likelihood concept is also included in Bayesian inference theory.

The ordering

$$(B) \Rightarrow (L) \Leftrightarrow (C) \Rightarrow (S)$$

of the four concepts indicated here (sufficiency, conditionality,

and likelihood as concepts of evidence; and Bayes' principle, denoted by (B)), is by their logical strengths; except that (C) and (L) are ordered by their evident intuitive and logical strengths as these seemed several years ago, before their equivalence was seen. (We tacitly assume at this point, as in previous papers, that the intuitively-trivial axiom (M') is adjoined to (C).) Throughout the writings of Fisher and Barnard, basic importance is ascribed to both (L) and (C), but with apparently different scopes and significance. Among those closer to the Neyman-Pearson theory, a time pattern of interest in, and tendency to accept, successively stronger concepts of evidence can be discerned.

One of the earliest applications of the Neyman-Pearson theory was the construction of binomial confidence intervals (the Clopper-Pearson charts, 1934). It seems impossible to account for the non-use of auxiliary randomization variables to obtain improved or exact results in the terms of the Neyman-Pearson theory, both here and in apparently all subsequent applications of that theory, except as an attempt to incorporate a sufficiency concept of evidence within the Neyman-Pearson theory. The comments of E. S. Pearson himself, discussed by Tukey (1962, pp. 12-13), seem to support this view strongly.

The appropriateness of a conditionality concept of evidence was stressed by Cox (1958) and Tukey (1958). And the equal appropriateness of a likelihood concept of evidence was seen to be entailed by the present writer's demonstration (1961, 1962) that (C) implies (L).

The pattern of these developments may be described as one of re-emergence of successive parts of the Bayesian concept of evidence on non-Bayesian ground. The neo-Bayesian writers (notably Savage (1954, 1961, 1962, 1954) and Good (1950)) have contributed significant new knowledge of possibilities of specification and interpretation of decision problems and related utilities and personal or subjective probabilities; and in this context a characterization of evidence coinciding formally with (L) appears. (An elementary expository version of the central neo-Bayesian result, the deduction of Bayes' principle from axioms characterizing rational decision-making behavior, has been given by Pratt et al (1964). A detailed analysis of the structure of Savage's axioms and derivation has been given by Morlat (1959).) But those questions which led to the decline and disuse of the classical Bayesian approach, the questions of specification and interpretation of prior probabilities in terms assimilable in the structure of empirical scientific work, remain unanswered. (This problem of giving adequate interpretation to prior probabilities is replaced by a "smaller" problem if prior information and opinion is assumed representable by a "prior likelihood" as described by Hudson (1964). For those inclined to accept the likelihood concept for experimental evidence, such as extension of the lieklihood concept would seem particularly attractive; but its value depends of course on an answer to questions of interpreting likelihood functions in general.)

The central result of these considerations is a trilemma concerning the concept of statistical evidence: The only theories

which are formally complete, and of adequate scope for treating statistical evidence and its interpretations in scientific research contexts, are Bayesian; but their crucial concept of prior probability remains without adequate interpretation in these contexts. Each of the non-Bayesian alternatives, one identified with the likelihood concept and the other with the error-probability concept, seems an essential part of any adequate concept of evidence, but each separately is seriously incomplete and inadequate; however these cannot be combined because they are incompatible (except, curiously, in the simplest restricted case, where one may say that a thoroughly satisfactory concept of empirical statistical evidence and its interpretation exist in miniature).

14. Acknowledgements.

The writer is indebted to Professors John Pratt, for helpful discussions of his "censoring" concept and its implications, and G. Morlat, for helpful discussions concerning Figure 2 and related interpretations.

Appendix: Further Details on the Axioms.

i. Pratt's strikingly simple, plausible, and original censoring concept seems to be the best elementary intuitive point of departure for appreciation of the major part of the likelihood axiom. An interesting alternative to Sections 2 and 3 above is one based on material from Sections 3, 8, and 10, presenting just (Ce), then (S), and then (L) and its equivalence to (Ce) and (S) jointly.

ii. The presentation of (H) in Section 11 stems from the writer's search for an intuitively-direct appreciation of Barnard's suggestions that the likelihood axiom, particularly when viewed as a concept of "irrelevance of stopping rules," is obvious. The method of proof of equivalence of (H) and (Ce) shows that there is no substantial mathematical distinction between sequential experiments and others: Each possible termination point of any experiment, sequential or not, can under relabeling be interpreted as a possible first-stage termination point of a sequential experiment, and so on. This is not at all incompatible with, but rather confirms, the *heuristic* value frequently found in sequential examples in discussions of concepts of evidence.

iii. The notation (C) might be reserved for the formulation of previous papers, which expresses the conditionality concept in its natural full scope, namely all mixture experiments. The notation (C') might be used for the weaker, simpler formulation which suffices in Section 9 above, where only mixtures of two components are considered. (Repeated applications of (C') extend its scope to mixtures of any finite number of components, but not to a countable infinity of components.) The main point of interest of course is just that even (C') (the formulation of this paper, called (C) here) suffices with (M') to imply (L); the slightly weaker assumptions make the result slightly stronger.

iv. The natural scope of the mathematical equivalence concept discussed in Section 7, all one-to-one transformations of S, might be denoted by (M) in distinction from the simpler, weaker (M')

which sufficed there. Previously (1962, pp. 277-8) this concept was described but not formulated axiomatically, and was tacitly assumed in assertions (unaccompanied by proofs) that (C) implies (S).

v. The natural scope of the sufficiency concept is all sufficient statistics, as in formulations denoted (S) in previous papers. The simpler, weaker formulation which suffices in Section 8 above might be given the different denotation (S'). (Repeated application of (S') gives an equivalent formulation covering all sufficient statistics in experiments with finite (but not countably-infinite) sample spaces.) It seems of pedagogical interest that the concepts of _statistic_ and _sufficient statistic_ were not introduced explicitly in Parts I and II above, although the material there includes what many would consider the principal significance of the sufficiency concept. Similarly presentation of (L) and its implications did not require explicit consideration of the likelihood function as a sufficient statistic.

References

1. Barnard, G. A. Review of Sequential Analysis by A. Wald. Journal of the American Statistical Assn. *42* (1947) 658-664.

2. Barnard G. A. Statistical inference. Journal of the Royal Statistical Society (B) *11* (1949) 116-149.

3. Barnard, G. A., Jenkins, G. M., and Winsten, C. B. Likelihood inference and time series. Journal of the Royal Statistical Society *A125* (1962) 321-327.

4. Birnbaum, Allan. On the foundations of statistical inference. I. Binary experiments. *Annals of Math. Statist*. *32* (1961) 414-435.

5. Birnbaum, A. On the foundations of statistical inference. *Journal of the American Statistical Association*, *57* (1962) 269-306, with discussion, pp. 307-326.

6. Clopper, C. J. and Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26* (1934) 404-413.

7. Cox, D. R. Some problems connected with statistical inference. *Annals of Mathematical Statistics* *29* (1958) 357-372.

8. Fisher, R. A. Theory of statistical estimation. *Proc. of the Cambridge Philosophical Soc*., *22* (1925) 700-725.

9. Fisher, R. A. *Statistical Methods and Scientific Inference*. Edinburgh (1956) Oliver and Boyd.

10. Good, I. J. *Probability and the Weighting of Evidence*. London: Charles Griffin, 1950.

11. Hudson, Derek. The use of a prior likelihood. Unpublished note (1964).

12. Morlat, G. Le choix d'une décision et le choix d'un critère. *La Décision* (1959) 123-128; Paris.

13. Nagel, Ernest. *The Structure of Science*. New York: Harcourt-Brace, 1961.

14. Neyman, J. and Pearson, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* *20A* (1928) 175- , 263- .

15. Neyman, J. *Lectures and Conferences on Mathematical Statistics*. Washington D. C.: Graduate School, U. S. Dept. of Agriculture, 1938 (2nd ed.: 1952).

16. Neyman, J. L'estimation statistique, traitée comme un problème classique de probabilité. *Actualités Scientifiques et Industrielles*. Paris: Hermann + Cie., pp. 25-57, 1938.

17. Neyman, J. Raisonnement inductif ou comportement inductif. *Proc. of the International Statistical Conf.*, *3* (1947) 423-433.

18. Neyman, J. 'Inductive behavior' as a basic concept of philosophy of science. *Review of the International Statistical Inst.*, *25* (1957) 7-22.

19. Neyman, J. Two breakthroughs in the theory of statistical decision making. *Review of the International Statistical Inst.*, *30*, No. 1 (1962) 11-27.

20. Pratt, John W. Review of *Testing Statistical Hypotheses* by E. L. Lehmann (Wiley, 1959), *Journal of the Amer. Stat. Assn.*, *56* (1961) 163-7.

21. Pratt, John W. Comments on A. Birnbaum's 'On the foundations of statistical inference,' *Jour. of the Amer. Stat. Assn.*, *57* (1962) 314-5.

22. Pratt, John W., Raiffa, Howard, and Schlaifer, Robert. The foundations of decision under uncertainty: an elementary exposition. Jour. of the Amer. Stat. Assn., 59 (1964) 353-375.

23. Robbins, H. Asymptotically subminimax solutions of compound statistical decision problems. Proc. of the Sec. Berkeley Symposium on Stat. and Prob., 1950, 131-148. Berkeley and Los Angeles· U. of Calif. Press, 1951.

24. Robbins, Herbert, and Samuel, Ester. Testing statistical hypotheses -- the 'compound' approach. Recent Developments in Information and Decision Processes, ed. by R. E. Machol and P. Gray, pp. 63-70. New York: Macmillan, 1962.

25. Savage, L. J. The Foundations of Statistics. New York: John Wiley + Sons, Inc., 1954.

26. Savage, L. J. The foundations of statistics reconsidered. Proc. of the 4th Berkeley Symposium on Math. Stat. and Prob., 1961.

27. Savage, L. J., et al. The Foundations of Statistical Inference. London: Methuen. New York: Wiley, 1962.

28. Savage, Leonard J. The Foundations of Statistics Reconsidered, Foundations of Subjective Prob., by Kyberg and Smokler. New York: Wiley, 1964. (Repr'd. with changes from 26. above.)

29. Smith, Cedric A. B. Consistency in statistical inference and decision (with discussion), Jour. of the Royal Stat. Soc. (B), 23 (1961) 1-37.

30. Stein, C. M. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proc. of the 3rd Berkeley Symposium on Stat. and Prob., 1 (1955) 197-206.

31. Stein, C. A remark on the likelihood principle. *Jour. of the Royal Statistical Society (A)*, 125 (1962) 565-573.

32. Tukey, J. W. *Fiducial Inference*. Unpublished: Wald Lectures, Cambridge, Mass., meeting of the Inst. of Math. Stat., 1958.

33. Tukey, J. W. The future of data analysis. *Annals of Math. Stat.*, 33 (1962) 1-67.

34. Walker, Helen M. and Lev, Joseph. *Statistical Inference*. New York: Holt, 1953.

35. Wilson, E. B. *An Introduction to Scientific Research*. New York: McGraw-Hill, 1952.